Pipeline Pilot Interface to *FTrees* Fragment Spaces

Version 2.4.5.1

User Guide

(for *FTrees* version 2.4.5 and above and Pipeline Pilot version 8.0 and above)

Edgar Derksen, Sally Hindle

Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen

The idea of Feature Trees was born in 1997 during Matthias Rarey's six-month research stay at SmithKline Beecham Pharmaceuticals R&D, King of Prussia, PA (USA) and then further developed at Institute for Algorithms and Scientific Computing (SCAI), then part of the German National Research Center for Information Technology (GMD) and now the Fraunhofer Gesellschaft (FhG). Since 2002 BioSolveIT GmbH has been responsible for the licensing and continuing development of the *FTrees* software.

At this point we would like to thank Scott Dixon (SmithKline Beecham, now Metaphorics LCC.), Markus Wagener (SmithKline Beecham, now N.V. Organon), and Jens Lösel for a lot of helpful and constructive discussions during Matthias Rarey's stay at Smithkline Beecham and afterwards. Without them, the idea of Feature Trees would not have been evolved in the way it has. Also, Matthias Rarey thanks the GMD and Smithkline Beecham for funding his research stay in King of Prussia.

In summer 2000, the Feature Tree comparison algorithms were extended to search directly in large combinatorial chemistry spaces. A two-stage dynamic programming algorithm enables searching directly in chemistry spaces without an explicit enumeration of molecules [2]. This work was also done during a research stay in the US, this time at Roche Bioscience in Palo Alto. The chemistry space search algorithm was developed in cooperation with Martin Stahl (Hoffmann-La Roche, Basel) and we would like to thank him for this excellent cooperation. We also wish to thank Hans-Joachim Böhm, Hans Maag (both Roche), and Thomas Lengauer (GMD) for making this research stay possible.

Since then the Feature Trees software has been further developed and extended by several contributors including Marc Zimmermann (FhG - MTrees and the new Dynamic Matchsearch algorithm), Robert Fischer, Sally Hindle, and other developers at BioSolveIT GmbH and the Center for Bioinformatics (ZBH), University of Hamburg.

This document contains proprietary information of BioSolveIT GmbH and is protected by copyright. It is provided together with Software of BioSolveIT under a license agreement and may be used only in accordance with the terms and conditions of this agreement. The document serves solely for the purpose of using the Software. No part of the document may be transferred to any third party or reproduced as a whole or in parts without written permission from BioSolveIT.

Base software: © 2001 by Fraunhofer Gesellschaft (FhI-SCAI); Getline library: © 1993 by Chris Thewalt; PVM library: © 1997 by University of Tennessee, Knoxville TN; Python library: © 1991-1995 by Stichting Mathematisch Centrum, Amsterdam, The Netherlands.

© 2015 BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany
Phone ++49-2241-2525-0, support@biosolveit.de

Contents

Contents **3**

1 Quick Start Steps **5**

1.1 Download and Import current *FTrees-FS* Package 5

1.2 Install *FTrees* 5

1.3 Create or Update Protocols 5

2 Introduction **7**

2.1 About *FTrees* 7

2.2 *FTrees-FS* Component in Pipeline Pilot 7

2.3 The Example Fragment Space: `tfs` 9

3 Installation **11**

3.1 BioSolveIT Web Installation 11

3.2 Using the *FTrees-FS* Installation Component 11

3.3 Using Global Variables 12

3.4 Using Custom External *FTrees* Installations 12

3.4.1 To Connect to an *FTrees* Installation on the Pipeline Pilot Server 13

3.4.2 To Connect to an *FTrees* Installation on a Remote Linux Server 14

3.5 Uninstallation 15

4 Trouble Shooting **17**

4.1 Problems connecting to the external installation 17

4.2 Problems using the *ssh* Method 18

4.3 Further help and BioSolveIT PDF Reporter 18

5 Tips and Tricks **21**

5.1 Other Significant Parameters in the *FTrees* Components 21

5.1.1 <Has Same File System> 21

5.2 Accessing Other Domains within Pipeline Pilot 21

References **23**

Bibliography **23**

1

Quick Start Steps

1.1 Download and Import current *FTrees-FS* Package

1. Download current "*FTrees-FS* in PipelinePilot" package on
   http://www.biosolveit.de/download/
2. Extract *FTrees-FS* package to a custom directory.
3. From the custom directory import all components (.xml) using PP-Client via drag and drop to your components collection.

1.2 Install *FTrees*

1. Drag and drop *FTrees* installation component from the components collection to an empty protocol.
2. Add your license for the executable of *FTrees* in the field for the parameter <Run FTrees -> on PP Server -> License Server or License File> in the Implementation tab – if your license is available from a license server, simply type in the name of the server in this format *@servername*, or if you have a license file then you may browse for it using the ... facility.
3. Run protocol once, it will result into a html report showing installation success.
4. For alternative installations and more details see chapter 3.

1.3 Create or Update Protocols

1. Drag and drop *FTrees* Fragment Space component into a protocol.
2. Update via a *right click drag and drop* of new component onto component reference in a protocol.
3. Now you are ready to setup the new components and run your protocol.

2

Introduction

2.1 About *FTrees*

FTrees is a piece of software for calculating the Feature Tree descriptor and comparing two or more of these descriptors to each other. The theory behind Feature Trees can be found in [1].

Rather than being based on a linear description such as bit strings or vectors, the Feature Tree descriptor represents the molecule as an unrooted tree where the nodes of the tree describe the major building blocks of the molecule. The comparison of two Feature Trees then proceeds using a recursive matching algorithm, splitting the trees into smaller and smaller subtrees. The Feature Tree approach has several advantages, the most important being the fact that the alignment of two Feature Trees can be translated into a comprehensible mapping of the two underlying molecules. For more details on the algorithms and achieved results see [1].

The very nature of the Feature Tree means it is also perfectly suited to working with fragments of molecules. Molecular fragments can be described by a Feature Tree just like a complete molecule and, therefore, can also be compared to each other. Together with the rules describing how fragments are allowed to be joined, the molecule fragments form a *FTrees* Fragment Space. Complete molecules can be used to search in the Fragment Space using the joining rules to guide the construction of fragments, see Figure 2.1 for an illustration. This technique is especially advantageous for searching in combinatorial sets of molecules like combinatorial libraries or combinatorial chemistry spaces. [2] contains more complete details about how the search algorithm works.

2.2 *FTrees-FS* Component in Pipeline Pilot

The *FTrees-FS* component can be thought of as a molecule generator for Pipeline Pilot. The component has no input port but takes a query file as a parameter. The results of searching with your queries is a set of resulting molecules output into the pipeline.

The component requires two files:

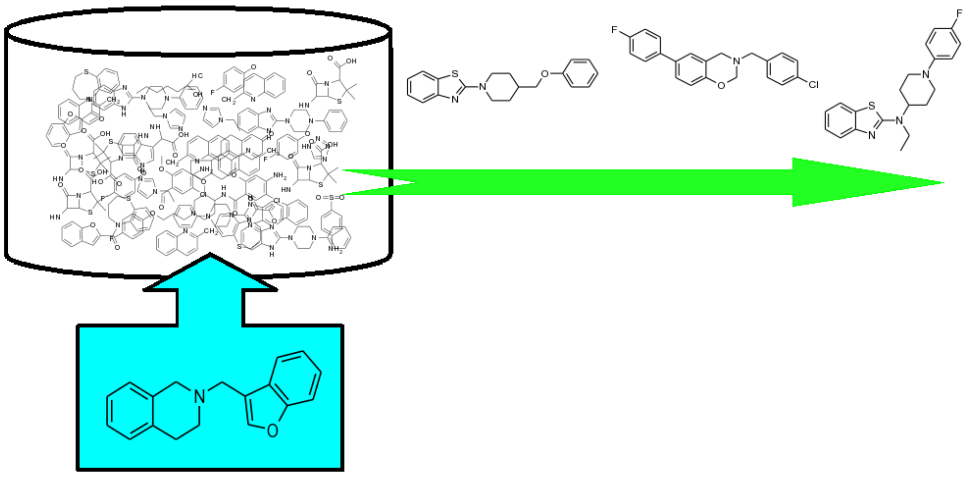

Figure 2.1: The *FTrees-FS* component takes a query molecule as input and uses it to generate new molecules from a Fragment Space based on a set of joining rules – the new molecules are sent via the Pass port into the pipeline

Fragment Space File A Fragment Space must be provided for the component as a `.zip` package. This is because a Fragment Space consists of several files and these *must* be kept together during any file transfer process. Inside the `.zip` package are the four files that form the Fragment Space in the format required by *FTrees-FS*. If you already have your own Fragment Spaces be sure to have all four files ready and pack them together using a zip program. The four files that form a Fragment Space in *FTrees-FS* format can be seen as an example in section 2.3.

Query The query file can be in any molecule input format readable by Pipeline Pilot, a Feature Tree file (`.fdf`) or the special Feature Tree database format used by the *FTrees* Writer and Reader components (`.ftdb`). You can use a multiple query file – make sure you adjust the maximum number of queries parameter.

The results of the search can be tuned according to how similar the resulting molecules are to the queries. Often, you do not want to see molecules almost exactly identical to your queries but are looking to generate new ideas. Change the parameter <FTrees Options -> Target Similarity> (an *FTrees* similarity value) to do this. If you find that the resulting molecules are too similar amongst each other, you can also set the diversity that must be seen between the resulting molecules. Change the parameter <FTrees Options -> Target Diversity>.

The resulting molecules exit the Pass port of the component. If you see any molecules exiting from the Fail port then these are queries that could not be interpreted correctly.

2.3 The Example Fragment Space: tfs

A ready to use Fragment Space, the knowledgespace, can be downloaded from http://www.biosolveit.de/download. The Fragment Space package knowledgespace-x.x.x-indep.zip must not be unpacked as the component will read the zip file directly.

Out of interest you may be curious about what is in the Fragment Space bundle. The following files are included:

knowledgespace.fsf This is the main Fragment Space file. It includes pointers to the molecule fragment file, Feature Tree fragment file, joining rule file and the joining rules themselves (linking rules)

knowledgespace.mol2 The actual molecule fragments themselves. The fragments were generated using publically available data.

knowledgespace.fdf The Feature Tree descriptors for the molecule fragments.

knowledgespace.ldf The linkage rules for the fragments.

3

Installation

3.1 BioSolveIT Web Installation

The easiest way to download and install BioSolveIT packages and tools is to download and run BioSolveIT Web Installer component. Download BioSolveIT in PipelinePilot package from http://www.biosolveit.de/download. Read UserGuide within that package for further details.

Note that both PP-Client and PP-Server need web access to run BioSolveIT Web Installer. Without web access you need to follow one of the following sections in this chapter.

3.2 Using the *FTrees-FS* Installation Component

The *FTrees* components are set by default to use a so-called auto installation of *FTrees*. This works as follows. When you run either of the *FTrees* Calculator or Similarity components (and the parameter <Run FTrees> is set to *on PP Server* and the parameter <Run FTrees-> on PP Server -> Use> is set to *FTrees Auto Installation*), they search for a *FTrees* installation in the directory:

```
<scitegic install directory>/public/bin/BioSolveIT/
```

and use this installation to run the calculation. A html report pupup shows if a *FTrees* installation is not found in this directory.

Use *FTrees* installer component to install *FTrees* prior to run calculations. In case of a default setup you don't need administrator rights to do this as it is actually Pipeline Pilot that carries out the installation and not you as a user. In special cases your system administrator prevents installations by PP and you need to ask him to install *FTrees* .

As this installation knows nothing about the licenses you may have for *FTrees*, you have to supply the license information separately. This is supplied using the parameter <Run FTrees -> on PP Server -> License Server or License File> as described above in the section 1.

The installation is carried out once only (for that *FTrees* version) and only once per server not per user. The software is available to all users once it is installed.

3.3 Using Global Variables

Use global variables as an alternative to the *FTrees* installation component decribed in section 3.2. Note that global variables will only be used if components are set to run auto installation (see. section 3.2) but no installation is done by an *FTrees* installation component yet.

Goto Administratin Portal accessible through the Server Home Page at http://localhost:9944/ if PP-Server is installed on your local machine. The Server Home Page is also available through your PP-Client in menu <Help -> Server Home Page>. Use default name: scitegicadmin and password: scitegic if you have not changed it yet. Browse to <Setup -> Globals>. Add Custom Global Protocol Properties.

Note that defining parameters within your protocol will override global variables if they are named the same. Use parameters within your protocol instead of global variables only if you want to use a custom installation for that protocol.

The first variable file_server_*FTrees*_executable: must be set to the full path of *FTrees* executable available by the PP-Server. Example:

```
C:\Program Files (x86)\BioSolveIT\FTrees\ftrees.exe
```

The second variable file_server_*FTrees*_config must be set to the full path of *FTrees* configuration. Example:

```
C:\Program Files (x86)\BioSolveIT\FTrees\config_ft.dat
```

In comparison to *FTrees* auto installation (see section 3.2) you need to check the compatibility between your defined installation and the components.

3.4 Using Custom External *FTrees* Installations

You can also use an existing external installation of the *FTrees* software at each component. This means you also have the opportunity to use settings different to those set by default. To do this, you must already have *FTrees* installed somewhere on your system outside of Pipeline Pilot. To install *FTrees* yourself, visit the download page at BioSolveIT:
`http://www.biosolveit.de/download`
and fetch the download package for your system for the latest *FTrees* package. Follow the instructions in the package to install *FTrees* and receive your licenses. Enter the license information for *FTrees* as described in the package **and not using the parameter <Run FTrees -> on PP Server -> License Server or License File> as for the internal installation**.

To use an external installation of *FTrees*, you must change the value of the parameter <Run FTrees-> on PP Server -> Use> in the Implementation tab to *preinstalled FTrees*.

There are actually two ways to use *FTrees* with an external installation. These are by using *FTrees* installed directly on the Pipeline Pilot server you are , or by accessing a remote machine where *FTrees* is installed using *ssh* method). The method is selected using the parameter <Run FTrees> Both methods are covered in more detail below.

3.4.1 To Connect to an *FTrees* Installation on the Pipeline Pilot Server

The most common scenario is that you will have an installation of *FTrees* on the Pipeline Pilot server. If you choose this option, you just have to enter the path to the executable and configuration file as parameters in the Implementation tab. Pipeline Pilot will then just start *FTrees* whenever it is required by making a call to the executable that you entered.

This method is given the name *on PP Server* – On Pipelinepilot Server. You can see an example in figure 3.1.

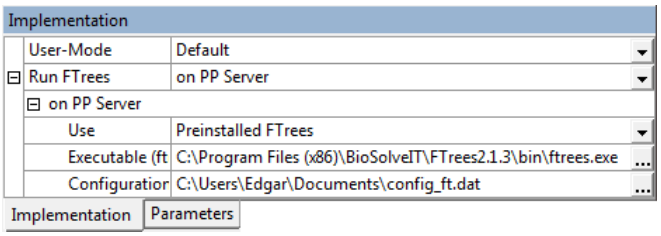

Figure 3.1: *ops*: Connect to an *FTrees* Installation on the Pipeline Pilot Server

- Requirement:
  - *FTrees* is installed on the Pipeline Pilot server. You can see which machine is the Pipeline Pilot server by starting your copy of Pipeline Pilot Client on your own workstation and find the name or IP of the server shown at the bottom right of the status bar (see figure 3.2). You must find where the *FTrees* installation is *on that machine*.

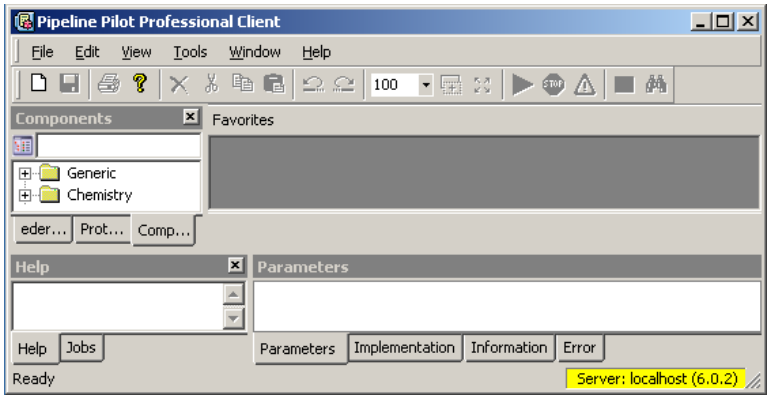

Figure 3.2: See where your Pipeline Pilot Server is installed.

- General Steps:
  1. Set <Run FTrees> to *on PP Server*
  2. Set <Run FTrees-> Use> to *Preinstalled FTrees*
  3. Expand <opsFTreesExe>

4. For the parameter <Run FTrees-> on PP Server -> Executable>, enter the path to the *FTrees* installation on Pipeline Pilot server. For example:
   `C:\Programs\BioSolveIT\FTrees2\bin\ftrees.exe`
5. For the parameter <Run FTrees-> on PP Server -> Configuration>, enter the path to the *FTrees* configuration file config_ft.datassociated with the *FTrees* installation. For example:
   `C:\Programs\BioSolveIT\FTrees2\config_ft.dat`

You can save these settings in the components.

3.4.2 To Connect to an *FTrees* Installation on a Remote Linux Server

Your existing *FTrees* installation could be on a Linux computer remote from the Pipeline Pilot server – in this case we offer an alternative so you can use the remote installation instead. Here, the calculations carried out by *FTrees* will be done on the remote Linux machine. Pipeline Pilot logs into the Linux machine using `ssh`, having copied all relevant files to the machine, and will run the calculation there, finally copying back all data it needs to the Pipeline Pilot server to continue with the pipeline.

This method is given the name *via SSH*. You can see an example in figure 3.3.

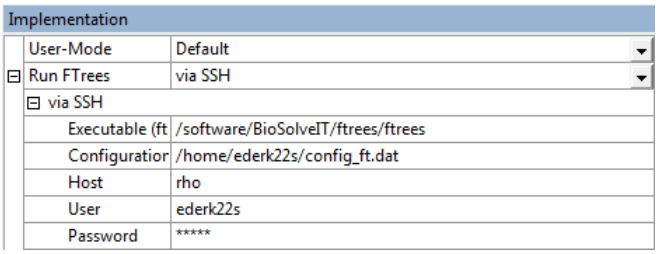

Figure 3.3: *via SSH*: Connect to an *FTrees* Installation on a Remote Linux Server via ssh

- Requirement:
  - *FTrees* is installed on a Linux machine available to the Pipeline Pilot server via `ssh`.
- General Steps:
  1. Set <Run FTrees> to *via SSH*
  2. For the parameter <Run FTrees-> via SSH -> Executable>, enter the path to the *FTrees* installation on the Linux machine. For example:
     `/software/BioSolveIT/ftrees-2.0.1/bin/ftrees-2.0.1`
  3. For the parameter <Run FTrees-> via SSH -> Configuration>, enter the path to the *FTrees* configuration file config_ft.datassociated with the *FTrees* installation. For example:
     `/software/BioSolveIT/ftrees-2.0.1/config_ft.dat`

4. For the parameter <Run FTrees-> via SSH -> Host>, enter the Linux machine host name

- User specific steps:

1. For the parameter <Run FTrees-> via SSH -> User>, enter the user login name for ssh on the Linux machine
2. For the parameter <Run FTrees-> via SSH -> Password>, enter the user password for ssh on the Linux machine
3. There are more advanced options to be found under <Run FTrees-> via SSH -> Options> for more specific ssh parameters. Note the option <Run FTrees-> via SSH -> Options -> Delete Results> may be useful for trouble-shooting later.

You can save these settings in the components (be sure not to save your own user specific login details in components available to others!).

All files necessary for the *FTrees* calculation will be transferred via `scp` between the Pipeline Pilot server and ssh Linux machine. Files copied and files created the remote server are automatically deleted at the end of the job leaving no trace. However, in case the user would like to leave a copy of the calculation and result files on the Linux machine, or for trouble-shooting as mentioned above, it is possible to set a parameter to tell Pipeline Pilot not to delete these files:
<Optionally> : <sshDeleteResults> : *False*

3.5 Uninstallation

Use the *FTrees* installer component to uninstall previous *FTrees* versions by switching Parameter "Uninstall Previous Versions" to "True".

Use BioSolveIT Uninstaller component to uninstall all *FTrees* versions.

Note that this kind of deinstallation only removes *FTrees* versions installed by *FTrees* installer component or BioSolveIT web installer component.

4

Trouble Shooting

4.1 Problems connecting to the external installation

Other most commonly seen problems with *FTrees* in Pipeline Pilot is with the connection to the external *FTrees* installation. For one thing, *FTrees* itself must be correctly installed on the system independently of Pipeline Pilot – it is essential first to make sure this is the case (especially to make sure that *FTrees* can locate the licenses). Once *FTrees* runs fine on your system, the remaining key task is to make sure the paths to the executable and configuration file are correct within the Pipeline Pilot components.

If something is amiss with the connection to the external installation, you will see an error message box pop up. Check the paths to the executable or to the configuration file. If it seems *FTrees* could be started but not run, the problem almost always lies either with the path to the configuration file or with licenses not being found. **Also make sure that you have a *FTrees* license** `[FFS]`**.** If *FTrees* runs OK independently from Pipeline Pilot then it is likely to be the path to the configuration.

When the error messages pop up, they may contain an FTreesError in the error message box, , as in figure 4.1.

Go to the 'Jobs' tab below the Protocol workspace and check under the last run job for a file called 'FTreesComponent Debug' – as in figure 4.2. Clicking on the link brings up HTML report with input and output data in a browser.

A correctly started *FTrees* job outputs the following header: if there is a problem you will see some of this header and the point where the problem occurs:

```
                 _____________________________________________________
               /                                                       /
             /                        F T R E E S                    /
           /        Feature-based molecular similarity finder       /
         /                                                         /
       /=========================================================/
     /                                                           /
   /  Version:  2.0.2  (11.08.08)                              /
 /  Modules: [PVM] [FFS]                                     /
/                                                           /
/ written by:  Matthias Rarey, Marc Zimmermann,           /
/              Sally Hindle, Robert Fischer             /
```

```
      /                                                          /
     / copyright by: BioSolve IT GmbH, FhG -SCAI               /
    /                Sankt Augustin, Germany                   /
   / for further information mail ftrees-info@biosolveit.de /
  /_____________________________________________________/
 Additional copyright notes:
   getline library: (C) 1993 by Chris Thewalt
   PVM library: (C) 1997 by University of Tennessee, Knoxville TN
 >> FTrees configuration file '/software/BioSolveIT/FTrees/ftrees2.0.1/
    config_ft.dat' loaded.

 >> Licensed modules: FTrees [PVM] [FFS]
 >> PVM status: no pvm daemon, running sequential.
 >> Scripts are executed in sequential mode; start PVM for parallel mode.
 >> SETTINGS = '/software/BioSolveIT/ftrees2.0.1/static_data/ftrees_settings.dat'
    loaded.
 >> CHEMPAR  = '/software/BioSolveIT/ftrees2.0.1/static_data/chempar.dat' loaded.
 >> CONTACT  = '/software/BioSolveIT/ftrees2.0.1/static_data/contact_ft.dat' loaded.
 >> TRANSFORM = '/software/BioSolveIT/ftrees2.0.1/static_data/transform.dat' loaded.
 >> GRAPHIC  = '/software/BioSolveIT/ftrees2.0.1/static_data/graphic_ft.dat' loaded.
...
```

4.2 Problems using the *ssh* Method

You may also experience problems using the *ssh* login, for example, the user name is unknown or the host is not found.

4.3 Further help and BioSolveIT PDF Reporter

More complicated errors may arise during the running of *FTrees*. Again though, the errors will be collected and as much information shown as possible. If you are familiar with *FTrees* you may want to take a look at all the output of the job yourself to see if you can recognize the problem. In this case, you can look in the temporary folders Pipeline Pilot sets up internally to find the output, or if you are working with the *ssh* method, set the parameter <Run FTrees -> via SSH -> Options -> Delete Results> to *False* so that you may then find the files retained on the *ssh* host: these will be in the directory set under the ssh parameter <Run FTrees -> via SSH -> Options -> Temp Path> (see the help text associated with this parameter to find its default value) – essentially a cryptically named folder whose name begins with the date and time of the job!

If you still do not know what is causing the errors, write down as much information as possible relating to your installation scheme. You can also create a PDF Report unsing BioSolveIT PDF Reporter component which summarizes installation data into one pdf file. Send all the information to:

support@biosolveit.de

Figure 4.1: An error box reporting that the *FTrees* exe could not be found.

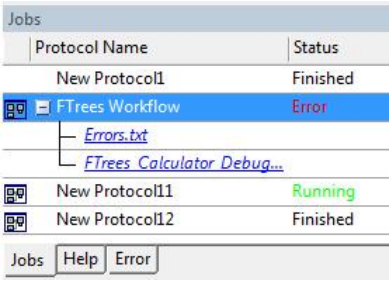

Figure 4.2: The full error report can be found in the 'Jobs' tab.

5

Tips and Tricks

5.1 Other Significant Parameters in the *FTrees* Components

For detailed documentation of all parameters for all components, refer to the documentation you find in the *Help* area of the Pipeline Pilot window.

We list here particularly interesting parameters: those that greatly influence the protocol or change the outcome of calculations, or those that may help you understand what is happening in the components.

5.1.1 <Has Same File System>

*FTrees*Calculator/Similarity : Implementation : <Run FTrees-> via SSH -> Options -> Has Same File System>

The name of this parameter stands for "ssh Has Same File System". Normally, for an *ssh* job, Pipeline Pilot must first copy all the data required by FTrees to the <Run FTrees -> via SSH -> Host> using scp. This is time consuming. It is possible that the Pipeline Pilot Server and <Run FTrees -> via SSH -> Host> actually share the same file system rendering the scp process unnecessary. Select True if the Pipeline Pilot Server and <Run FTrees -> via SSH -> Host> share the same File System - no copying of data is necessary. Selecting False means Pipeline Pilot copies all the data to and back from the <Run FTrees -> via SSH -> Host> – this is just a little slower but will always still work. Leave the parameter set to False if you are uncertain!

5.2 Accessing Other Domains within Pipeline Pilot

Often in house data or even your own working data are accessible from a windows computer via a domain (a path starting for example 'z:\...' or '\\...') which you cannot find from within Pipeline Pilot. That means you must first literally transfer the data to the Pipeline Pilot Server itself.

If you are using a Linux Pipeline Pilot server, this hint does not apply.

To get around this problem and make the Pipeline Pilot working environment much more flexible you can allow users access to domains – you need Administrator rights to be able to do this! Also check first that you should change these settings as they may have already been set to fit the current environment.

- Go to the 'Scitegic Server Home Page', for example, via the Help menu in your Pipeline Pilot client.
- Click on 'Pipeline Pilot Administration Portal' and log in with the Administrator user name and password.
- In the Security tab go to Authentification.
- For the 'Authentification Method' choose 'DOMAIN' and a set of parameters will appear.
- Enter the domain name in the field 'Domain' and choose 'Full' for 'Impersonation'
- Choose 'DOMAIN' for 'Retrieve Groups' and leave 'Limit access to listed domains' set to 'No'
- click 'Save' and log out again.

After you have done this you will need to enter your domain login details when you start the Pipeline Pilot Client.

Bibliography

[1] M. Rarey and J.S. Dixon. Feature trees: A new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design*, 12:471–490, 1998. 7

[2] M. Rarey and M. Stahl. Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design*, 15:497–520, 2001. 2, 7